

**Tax Policy and the Missing Middle:
Optimal Tax Remittance with Firm-Level Administrative Costs**

**Dhammika Dharmapala
University of Connecticut**

**Joel Slemrod
University of Michigan**

**John Douglas Wilson
Michigan State University**

**This version:
April 11, 2008**

We are grateful for comments on an earlier draft received at seminar presentations at Hunter College, Queen's University, Stanford University, the University of California at Berkeley, and the World Bank.

1. Introduction

The phenomenon of the “missing middle”, in which many small firms and a few large firms produce the bulk of value added, is ubiquitous in developing countries. It forms part of a wider cluster of interrelated characteristics – including a large informal sector and regulatory barriers to entry into the formal sector – that have attracted increasing attention in the literature. These features have often been explained with reference to the “grabbing hand” associated with bureaucratic inefficiency and corruption (e.g. Friedman *et al.*, 2000; Djankov *et al.*, 2002). In the development literature it has been suggested that the missing middle arises in part because taxes and regulations are enforced only among large, formal-sector firms (e.g. Rauch, 1991).¹ Notably, though, the development literature does not address under what conditions it is *optimal* for policy to treat differentially firms of different size, perhaps in a way that generates a missing middle. Nor, indeed, has the public finance literature focused on this issue, largely because its standard normative framework does not allow for heterogeneous firm size in a meaningful way.

In this paper, we provide a normative framework for analyzing policies that apply differentially to small and large firms, and demonstrate that under some conditions optimal policies will generate a missing middle — a range of output that no firm produces. We address tax policy, although some of the insights are also relevant to regulatory policy. Our basic argument is that the government should economize on administrative costs by exempting small firms from taxation, even though doing so causes intermediate-sized firms to reduce their outputs to tax-exempt levels, thereby creating a missing middle. The approach is consistent with Gordon and Li’s (2005) basic contention that the policies pursued by developing countries can potentially be explained on normative grounds, rather than through political economy explanations based on the self-interest of non-benevolent policymakers. More broadly, a central aim of our paper is to construct a model of optimal tax policy that recognizes the central role that firms play in the remittance systems of all modern taxes, and the importance of treating firms differently according to their size, regardless of whether these taxes are levied in developing or developed countries.

¹ Tybout (2000) ably reviews the empirical evidence about this phenomenon.

The importance of firms becomes apparent once one recognizes that it is cost-efficient for the tax authority to deal with a small number of entities with relatively sophisticated accounting and financial expertise rather than a much larger number of employees or providers of capital. The centrality of firms is illustrated by two recent studies that find that over 80 percent of all taxes are remitted by business in the U.S. and the U.K.² Anecdotal evidence suggests that collection of taxes from businesses is even more important in developing countries.

Notably, though, dealing with *small* businesses is not generally cost-efficient, and many tax systems either entirely exempt small businesses from remittance responsibility, or else feature special tax regimes for small businesses that simplify the tax compliance process, and thereby change the base on which tax liability is based. In many countries the exemption of small firms is *de facto*, due to lax enforcement.³ One implication of these policies is that the collection of taxes is highly concentrated among relatively large firms. A recent report asserts that the typical distribution of tax collections by firm size for African and Mid-Eastern countries would be that less than one percent of taxpayers remit over 70 percent of revenues, and the report gives specific examples of highly concentrated patterns: in Argentina, 0.1 percent of enterprise taxpayers remit 49 percent of revenues; and in Kenya, 0.4 percent remit 61 percent.⁴

Although explicit or implicit exemption, or more generally special tax treatment, of small firms might economize on collection costs (both compliance costs borne in the first instance by taxpayers and administrative costs borne in the first instance by the tax authority), it also generally causes production inefficiency, in part because it provides a tax-related incentive for firms to be — or stay — small. The tradeoff between the costs of collection and production inefficiency has not been closely addressed by the optimal tax literature. In part, this is because nearly all of modern tax theory is concerned with what actions or states of the world *trigger* tax liability, and virtually none is concerned with the system of remittance of funds to the government to cover that liability. Indeed, elementary public finance textbooks often assert that the remittance details — such as

² Christensen, Cline, and Neubig (2001) and Slemrod, Whiting, and Shaw (2007).

³ For an excellent review of the sorts of special regimes that countries apply to small businesses, see International Tax Dialogue (2007).

⁴ The report is International Tax Dialogue (2007). Gauthier and Gersovitz (1997) document this phenomenon for Cameroon.

whether the buyer or seller of a commodity remits the sales tax triggered by the sale — are *irrelevant* to the consequences of a tax.⁵ The operation of actual tax systems, however, requires considerable attention to the remittance of monies to the tax authority, including both the administration and enforcement of the tax rules, as well as the design of the tax rules with the administrative and compliance issues in mind. This is especially true in developing countries, where administrative constraints are often first-order issues in tax systems.

In addition, the famous Diamond and Mirrlees (1971) theorem on aggregate production efficiency demonstrates that production inefficiencies should not be tolerated if the government faces no constraints on its ability to levy optimal commodity taxes. But their model of optimal taxation ignores administrative costs and assumes that there are no untaxed profits, due either to constant returns to scale or a 100 percent profits tax. Yitzhaki (1979) and Wilson (1989) investigate optimal commodity taxation when there is costly tax administration, but their assumption of constant returns to scale eliminates any role for firms. An earlier paper by Heller and Shell (1971) presents a general framework for analyzing an optimal system of commodity taxes, lump-sum taxes, and firm-specific licensing fees and profits taxes when these tax instruments are costly to administer. In contrast, a major goal of the present paper is to investigate the optimal taxation of individual firms when firm-specific taxes are not available and therefore exemptions from taxation must be based on observed outputs. Keen and Mintz (2004) consider an output cutoff for exempting firms from a value-added tax in the presence of administrative and compliance costs. However, firms expect to earn untaxed profits in their model, in which case the Diamond-Mirrlees theorem no longer holds and different firms should generally be taxed at different rates, even without administrative costs.⁶ As we next explain, administrative costs are solely responsible for the production inefficiencies in our analysis, and these costs require an expansion in the set of tax instruments.

⁵ The relevance of the remittance system is discussed in Slemrod (forthcoming).

⁶ This discussion refers to the second of two models presented by Keen and Mintz (2004). In the first model, firm sales are exogenous, so a cutoff rule does not distort output decisions. (See also Zee (2005) for an extension of this model.) Both models differ from our analysis by treating net product prices as fixed (a small open economy), assuming an exogenous social marginal value of government revenue, and not allowing a per-firm fixed fee as a policy instrument.

In our main model, the three available policy instruments are a constant tax rate on output, a fixed per-firm fee, and an output cutoff, below which firms are not taxed. While entry fees have often been viewed as a manifestation of bureaucratic inefficiency or corruption (e.g. Djankov *et al.*, 2002), we show that when all firms in an industry are taxed, optimal policy may involve the use of the fixed fee;⁷ the fee basically acts like a Pigouvian tax, internalizing the social costs of tax administration.^{8,9} In our model, each *industry* is characterized by constant returns to scale, because the set of firms that are potential producers is effectively unlimited and *ex ante* identical. We focus on models in which firms “discover” heterogeneous productivities after entering the industry.¹⁰ In this setting, the standard rules of optimal commodity taxation hold if there are no administrative costs, enabling us to isolate the implications of introducing these costs. In particular, the Diamond and Mirrlees (1971) theorem on aggregate production efficiency tells us that the tax system should not discriminate among firms in the same industry. With administrative costs, we identify cases in which it is optimal to exempt small firms from taxation. This creates production inefficiencies that are inconsistent with the optimal tax system in the Diamond and Mirrlees framework. These inefficiencies occur because different firms in the same industry sell output at different prices, and also because some firms obtain the tax exemption by reducing their outputs to inefficiently low levels, creating the “missing middle” described above.

As a precursor to our model of heterogeneous firms, we first investigate the optimal system of output taxes and fixed fees in an economy where firms differ only across industries, not within industries. In particular, we characterize the optimal system of output taxes and fixed fees, and we derive a rule to determine when administrative costs are sufficiently high to justify the exemption of an industry from taxation. After

⁷ Auriol and Warlters (2005) also argue that entry barriers may be optimal in some circumstances, but their mechanism is very different: the entry barriers generate rents for incumbents that the government can tax.

⁸ International Tax Dialogue (2007, p. 31) discusses the fixed per-firm fee, also known as a *patente* system, as an example of a presumptive tax regime. The fixed per-firm fee could also be motivated by the entry fees and registration costs imposed by governments on firms, as measured and analyzed by Djankov *et al.* (2002).

⁹ If, instead of administrative costs, firms incurred fixed compliance costs in the payment of taxes, then the fee would not be needed because these costs would already be internal to the firm. Our other results are also easily modified to account for compliance costs.

¹⁰ The earlier literature allows for firm heterogeneity (e.g. Fortin, Marceau and Savard, 1997), but typically involves *ex ante* exogenous differences among firms, whereas our approach generates endogenous firm heterogeneity.

obtaining these results in Section 2, we devote Section 3 to the development of the heterogeneous-firm model. In Section 4, we show that under certain conditions, cutoffs are part of the optimal tax structure. In Section 5, we introduce an optimal nonlinear tax on an industry's taxed output, emphasizing the use of the nonlinear structure to reduce taxes on firms that might otherwise avoid these taxes by inefficiently reducing their outputs to the cutoff level. In Section 6, we develop a modified inverse-elasticity rule for the average net (of administrative costs) taxes on different goods. Section 7 concludes.

2. The Structure of Firm-Remitted Taxes across Homogeneous Industries

We begin by focusing on differences across industries rather than on differences among firms within an industry. To do so, consider an economy with many industries, each of which has access to an unbounded mass of identical potential firms. Before making its entry decision, a firm knows that if it enters an industry, it will face a strictly increasing, strictly convex cost function given by $c_i(y_i) + c_{ei}$ for a firm with output y_i in industry i , where the derivative, $c_i'(y_i)$, is positive at all y_i , and c_{ei} is treated as an entry cost. When we later introduce heterogeneous firms, we will assume an additional fixed production cost incurred by firms after entering and choosing to produce. In either case, the entry cost and increasing marginal cost imply U-shaped average cost curves.

The tax instruments consist of a constant marginal tax rate on output, t_i , and a fixed fee, b_i . The fixed fee will turn out to be a critical component of the tax system when there are per-firm administrative costs in the collection of taxes. Letting p_i denote the producer price for good i , calculated net of the output tax, profit maximization yields a firm's output function, $y_i(p_i)$, and its profit function, defined gross of the fixed entry cost:

$$\pi_i(p_i, b_i) = p_i y_i(p_i) - c_i(y_i(p_i)) - b_i. \quad (1)$$

Free entry guarantees that p_i will settle at where $\pi_i(p_i, b_i)$ equals the entry cost.

The demand for each good is obtained from the utility-maximization problem for a representative consumer with utility function, $U(X, L)$, where X is a vector of I goods and L is the supply of an input called "labor." The representative consumer supplies this labor to firms in each industry and receives all profits, which are zero in equilibrium.

Labor serves as the *numeraire*, so the costs described above are measured in units of labor. So that we can work with demand functions for each good that depend only on the good's own price, we assume that the direct utility function is separable and quasi-linear in labor. Letting q_i denote the price that the consumer pays for good i , utility maximization gives the demand functions, $X_i(q_i)$ for good i , and the indirect utility function, $\Sigma_i v_i(q_i)$. In the presence of a unit tax t_i on good i , the consumer price q_i equals $p_i + t_i$. The number of firms producing good i , M_i , is determined by the requirement that demand equals supply: $X_i(q_i) = M_i y_i(p_i)$.

In this paper we focus on the implications for optimal tax policy of a fixed per-firm cost of collecting taxes.¹¹ Specifically, the government incurs a fixed per-firm “administrative cost”, A_i , when it collects taxes from a firm. This assumption is intended to capture the notion that there is a substantial fixed component to tax administration and enforcement. For instance, suppose that tax enforcement requires that (at least with some positive probability) the tax authority must dispatch agents to audit the records of each firm. Although it may be the case that auditing a larger firm requires more resources, it is exceedingly unlikely that such differences, even if they exist, will be proportional in size to the differences in firms' output levels.¹² Although to simplify the model we assume that administrative costs are fixed and hence independent of the size of the firm, the qualitative results apply as long as there is a relatively large fixed component to these costs.

The government's tax instruments consist of the vectors of output taxes and fixed fees. The values of these control variables determine the market-clearing values of the consumer and producer prices, and the number of firms entering each industry. Following the standard practice in optimal tax theory, it is convenient to take as control variables the consumer and producer prices, rather than the tax rates.¹³ It is also convenient to make the M_i 's control variables.

The Lagrangian describing the government's problem is as follows:

¹¹ We do not allow firms to split up or combine for tax purposes.

¹² See Slemrod (2006) for a review of the empirical evidence showing that the *compliance* costs of business taxes as a fraction of firm size fall with firm size, suggesting a significant fixed aspect of this part of tax collection costs.

¹³ In standard constant-cost models, p_i is fixed by the technology, but we shall see that this is not the case here.

$$L = \sum_i v_i(q_i) + \sum_i \lambda_i (X_i(q) - M_i y_i(p_i)) \quad (2)$$

$$+ \lambda_B \left(\sum_i (M_i ((q_i - p_i) y_i(p_i) + b_i - A_i)) - E \right) + \sum_i \beta_i (\pi_i(p_i, b_i) - c_{ei}).$$

The government seeks to maximize the utility of the representative consumer, subject to three types of constraints. The constraint multiplying the Lagrange multiplier λ_i is the requirement that demand equal supply in the market for good i . The multiplier λ_B applies to the government budget constraint, where E is an exogenous revenue requirement. Finally, each multiplier β_i pertains to the requirement that profits equal zero in industry i .

We first show that the fixed fee should equal the per-firm administrative costs, leaving the output tax to finance expenditure needs. (All proofs appear in the Appendix.)

Proposition 1. *If firms producing good i are taxed, then the optimal fixed fee equals administrative costs: $b_i = A_i$.*

Proposition 1 can be understood as establishing that the optimal value of b_i is a Pigouvian tax: a taxed firm generates a social cost in the form of administrative costs, and the fee should internalize this cost.¹⁴ This reasoning clearly does not depend on the simplifying assumption that all firms are identical, so Proposition 1 carries over to the model of heterogeneous firms addressed beginning in the next section, assuming all firms are taxed. Note that, *ceteris paribus*, the higher is A_i (and therefore the optimal value of b_i), the larger is the optimal output of a firm in industry i , and the lower is the optimal number of such firms. This is true because a higher b_i requires a higher p_i to maintain the zero-profit condition, which in turn implies a higher y_i ; for given q_i and thus total output $X_i(q_i)$, this in turn implies a smaller M_i . This zero-profit condition implies that each firm produces at the bottom of its average cost curve, inclusive of b_i . Proposition 1 then tells us that average production and administrative costs are minimized.

With the fee addressing the external nature of the administrative costs, we should expect the output tax to satisfy the usual inverse-elasticity rule for an optimal tax

¹⁴ See Slemrod (2008) for a discussion of the extent to which administrative and compliance costs are appropriately treated as externalities, and under what circumstances the optimal tax treatment follows the targeting principle of Sandmo (1975) and generalized by Kopczuk (2003).

system.¹⁵ This turns out to be the case. Let α denote the marginal utility of income, and define the price elasticity of demand, measured positively, as follows:

$$\varepsilon_i^X = -\frac{dX_i}{dq_i} \frac{q_i}{X_i}. \quad (3)$$

We then have:

Proposition 2. *If all firms producing good i are taxed, then the optimal output tax structure satisfies the inverse-elasticity rule:*

$$\frac{t_i}{q_i} = \frac{1 - \frac{\alpha}{\lambda_B}}{\varepsilon_i^X}. \quad (4)$$

Proposition 2 asserts that, as b_i addresses the externality, any remaining net revenue requirement ought to be collected as would otherwise be optimal, in this case by following the well-known inverse elasticity rule.

In the case of heterogeneous firms considered in the next section, the set of producing firms and the set of taxed firms may be differentiated by establishing a size threshold for being subject to tax. But under the current assumption of homogeneous firms, producing firms can be distinguished from taxed firms only by exempting entire industries from taxation. Exemption of an industry might be optimal because, although administrative costs do not affect the optimal structure of output taxes on taxed goods (by Prop. 2), they may affect the optimal set of taxed goods.

Suppose, in particular, that for some reason administrative costs A_i incurred in taxing firms in sector i increase, while the costs related to another sector j (or set of other sectors) decrease, so that the government budget stays balanced with no change in taxes. Then (4) tells us that the optimal structure of output taxes does not change, and therefore social welfare stays constant, *if* we continue to tax the same goods. Instead, the government finances the higher A_i by raising the fixed fee b_i , and lowering b_j . If, however, A_i gets sufficiently high, then the government should reason that, if the initial

¹⁵ See Auerbach and Hines (2002) for a careful exposition of this model.

output taxes are low relative to administrative costs, and a good is not such a large contributor to the government budget that exempting it would require much higher taxes on the remaining goods, then it should be exempted from taxation altogether in order to obtain the savings in administrative costs. Using the quadratic formula for the deadweight loss of a tax, we next state a sufficient condition for industry exemption.¹⁶

Proposition 3. *Assume that all firms in all industries are initially taxed according to the inverse-elasticity rule, and no industry is large enough to generate more than a small share of total tax revenue. Then social welfare can be increased by exempting good j from taxation if the ratio of total administrative costs to total output tax payments for good j satisfies*

$$\frac{M_j A_j}{t_j X_j(q_j)} > \left(\frac{2MCPF - 1}{3MCPF - 1} \right) (MCPF - 1), \quad (5)$$

where $MCPF$ is the marginal cost of public funds and linear approximations are used to measure the effects of taxes on demands.

Thus, it is desirable to exempt an industry from taxation if the ratio of administrative costs to tax revenue exceeds a lower bound. It is interesting that this lower bound is less than the deadweight loss from another dollar of tax revenue, measured by the amount by which the marginal cost of public funds exceeds one. The

Appendix shows that this marginal cost is $MCPF = \frac{1}{1 - \frac{t_i}{q_i} \varepsilon_i^X}$, but (5) multiplies the

excess of this cost over one by a number less than one. The proof of this result is complex, but some rough intuition can be provided. If we eliminate taxes on good j , then we must raise taxes on other goods to offset the revenue loss. The welfare loss from these other changes depends on the marginal cost of public funds, so this marginal cost appears on the right side of (5). However, eliminating taxes on j also gives us an

¹⁶ Condition (5) is a sufficient condition, but it is not a necessary condition because it ignores some of the gains in consumers' and producers' surplus that arise from the elimination of the fixed fee; see the proof in the Appendix.

efficiency gain, which depends on the term, $\frac{t_i}{q_i} \varepsilon_i^X$, in the expression for the *MCPF*.

This reasoning leads us to expect the relatively small lower bound in (5).

For an illustrative example, note that although there is a wide range of empirical estimates of the *MCPF*, 1.25 is a reasonable midpoint of the estimates.¹⁷ This value implies that $\frac{t_i}{q_i} \varepsilon_i^X = 0.2$, which would hold, for an example, with an elasticity of demand

equal to one and a twenty percent tax rate.¹⁸ With this value for the marginal cost of public funds, condition (5) becomes

$$\frac{M_j A_j}{t_j X_j} > .136. \quad (6)$$

This implies that it will be desirable to completely exempt a good from taxation if its administrative costs are more than about fourteen percent of output tax revenue. This criterion could plausibly be satisfied for industries consisting of relatively small firms.

Proposition 3 has interesting parallels to the work of Yitzhaki (1979) and Wilson (1989), both of which consider the optimal taxation of a continuum of goods that enter a representative consumer's utility function symmetrically.¹⁹ In these models, taxing any good incurs an administrative cost that varies across goods, which suggests that taxing fewer goods is better than taxing many goods. But expanding the tax base reduces the standard deadweight loss from taxation and, at the optimal number of taxed goods (tax base breadth), the marginal administrative cost equals the marginal saving in deadweight loss. The current model addresses a similar tradeoff, and could be used to analyze the optimal number of taxed goods in an economy by applying the logic of Proposition 3. As the number of untaxed goods rises, the increased taxes on the remaining taxed goods would raise the marginal efficiency losses from taxing them.

In the Yitzhaki and Wilson models, however, the reason that administrative costs vary across goods is exogenous. Nor could the source of variation be related to

¹⁷ See Ballard (1997).

¹⁸ If we interpret this tax rate as the one that would prevail when all goods are taxed, then it should be viewed as lying below the tax rates under the tax structures that countries typically use, because they tend to exempt many commodities from taxation.

¹⁹ See also Slemrod and Kopczuk (2002), which adds distributional concerns to the choice of how broad the tax base should be.

administrative costs at the firm level, as the models adopt the standard assumption that all firms exhibit constant returns to scale, rendering the size of firms indeterminate. In the present model, U-shaped average cost curves limit the equilibrium size of firms. Heterogeneity of production technologies introduces heterogeneity across sectors in the cost of collecting taxes, as it is more costly to collect taxes from industries whose technology favors small firms, even when the tax authority optimally uses the policy instruments at its disposal (including both the fixed fee that mirrors the fixed per-firm administrative costs and, as modeled below, a threshold size for being subject to tax). In this manner, we endogenize inter-industry differences in administrative costs and obtain a tradeoff between these costs and the deadweight cost of taxation that is similar to the tradeoff studied by Yitzhaki and Wilson. As evidenced by Proposition 3, industries with many small firms are likely to exhibit relatively high administrative costs in tax collection, increasing the likelihood that they should be exempted from taxation. Consequently, the tax system favors small firms, while penalizing large firms through higher tax rates. Eventually, these higher tax rates will cause the sufficient condition (5) to be violated, limiting the number of industries that it is optimal to exempt.

3. A Model with Heterogeneous Firms

Within an industry, firms typically differ significantly in size. For example, small grocery stores and large supermarkets both sell food. Given our assumption that administrative costs have an important fixed per-firm component, it follows that it might be optimal to exempt small firms in an industry from taxation. As previously noted, the explicit or *de facto* exemption of small firms is a widespread phenomenon in tax systems. Thus it is especially important to have a theoretical framework that will enable the rigorous analysis of the optimal structure of such policies.

The model we develop is inspired by Hopenhayn (1992a, b) and Melitz (2003), but has been simplified to de-emphasize the dynamics that are a central focus of those papers. It makes use of the competitive setting used by Hopenhayn (1992a, b), rather than the monopolistic competition model of Melitz (2003).

We first describe the behavior of firms, and then turn to the government's optimal tax problem. Because we are interested in how firms within a given industry should be

taxed, at first we drop subscripts identifying goods and focus on a single industry. In particular, we assume that the government has chosen the consumer price for this industry, q , fixing demand at $X(q)$, and we solve the sub-optimization problem of maximizing net revenue, given q . If revenue were not maximized, then it would be possible to move to a different tax system that created a budget surplus that could be passed on to consumers through a welfare-improving reduction in q . In Section 6 we solve the more general problem that includes choosing the price vector faced by consumers, and in so doing derive a modified inverse-elasticity rule for how the consumer prices on different goods should be chosen.

Building on our previous model, assume again that all firms in the industry face the same fixed cost, c_e , but now allow that the convex variable cost function differs across firms: $c(y, \varphi)$ for a type- φ firm, where φ is an *ex ante* unknown productivity parameter that takes on values over an interval, $[\varphi^l, \varphi^h]$.²⁰ The value of φ cannot be discovered (by the entrepreneur) unless the firm enters, although its distribution (characterized by cdf $F(\varphi)$ and pdf $f(\varphi)$) is known *ex ante*, and there is an unbounded mass of identical potential entrants, each possessing this distribution.

Each firm chooses its output only after incurring the fixed cost, c_e . We assume that a higher value of φ decreases marginal costs: $\frac{\partial c(y, \varphi')}{\partial y} < \frac{\partial c(y, \varphi'')}{\partial y}$ for $\varphi' > \varphi''$ and all y . Thus, a firm's chosen output rises with φ . In addition, we assume that $c(y, \varphi)$ is bounded from below by a positive number at all positive y , reflecting a fixed cost of production, which we assume for simplicity to be constant across firms.

For a firm facing unit output price p and fixed fee b , profits, calculated gross of the fixed entry cost, are given by

$$\pi(p, b, \varphi) = py(p, \varphi) - c(y(p, \varphi), \varphi) - b, \quad (7)$$

where the supply function, $y(p, \varphi)$, is obtained by maximizing profits. These profits are increasing in φ .

²⁰ Unless otherwise stated, φ^l and φ^h can be taken to be minus infinity and plus infinity, respectively.

Once a firm has entered the industry, it incurs no additional costs if it exits the industry before producing (i.e., $c(y, \varphi)$ discontinuously drops to zero at $y = 0$). Thus, it will exit if the φ it draws is too small to yield non-negative profits; that is, if $\pi(p, b, \varphi) < 0$. Although exit entails no costs, the entry cost c_e is sunk and cannot be recovered. For the subsequent analysis, we assume that there always exist some firms whose productivity is low enough that they exit.

Now consider introducing a tax-threshold level of output, y^* , such that firms with output equal to or less than y^* are not taxed and thus receive the consumer price q per unit of output. Firms with output above y^* pay the fixed fee b and are taxed at the rate t on their output, in which case they receive the producer price $p (= q - t)$ per unit of output. The assumption here is that the government is unable to observe productivities directly, and must therefore base its tax on observed outputs.

If y^* is sufficiently low, then no firm will be willing to produce untaxed output, because a firm cannot cover their fixed costs if its output is too low. As y^* increases, however, eventually some firms producing above y^* will be tempted to lower their output to y^* to escape taxation. Higher levels of y^* can induce some firms to choose outputs below y^* . But there will still be firms that produce at y^* , but would produce above y^* if the tax break at y^* were not available.

Figure 1 illustrates the incentives facing two such firms, which differ in their productivities. The optimal taxed outputs for these high- and low-productivity firms are located where their marginal cost curves, MC^h and MC^l , equal the producer price p . The high-productivity firm incurs a loss in producers' surplus of area $a+b+c$ from the drop in output from y_h to y^* , but this drop is offset by the elimination of the tax burden once y^* is reached. The low-productivity firm incurs a smaller loss in producers' surplus, given by area a , because its marginal cost curve is higher by assumption. Thus, it too reduces output to y^* . There is then bunching at output y^* of firms with productivities within some interval. Moreover, this bunching eliminates the production of outputs between y^* and some higher output, giving us the "missing middle" proposition discussed in the introduction:

Proposition 4 (the “missing middle”). Under an optimal linear tax system with a cutoff y^* that induces some firms not to pay taxes, there exists a $y^{**} > y^*$ such that firms with sufficiently high productivities produce outputs above y^{**} , but no firms produce an output between y^* and y^{**} .

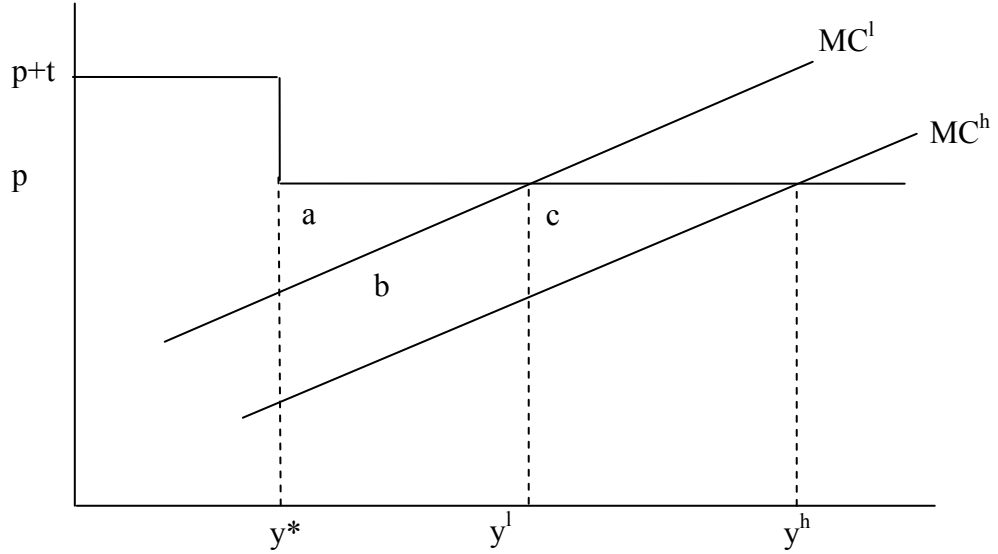


Figure 1

Thus, the economy contains small firms and large firms, but is missing firms with intermediate levels of output. Firms that would be producing these outputs instead reduce their outputs to y^* to eliminate their tax burdens. The lowest taxed output that firms are willing to produce, y^{**} in Proposition 4, leaves the marginal firm indifferent between y^* and y^{**} :

$$qy^* - c(y^*, \varphi) = \pi(p, b, \varphi). \quad (8)$$

Solving this equality for φ gives us the lowest productivity possessed by firms producing taxable output, defined as a function of prices, y^* , and the fixed fee b : $\varphi^{**}(q, p, b, y^*)$.

With this notation, the total output of taxed firms is

$$Y^T = M \left(\int_{\varphi^{**}(q, p, b, y^*)}^{\varphi^h} y(p, \varphi) f(\varphi) d\varphi \right). \quad (9)$$

where M is, as before, the number of firms entering the industry.

Consider now the determination of untaxed output. If y^* is not too high, then there will be no firms producing below y^* . The lowest productivity of firms that do produce y^* will satisfy the zero-profit requirement, $qy^* - c(y^*, \varphi) = 0$, allowing us to define this lowest productivity as a function of q and y^* : $\varphi^m(q, y^*)$. Higher values of y^* enable some firms to maximize profits at untaxed outputs below y^* . We may still let $\varphi^m(q, y^*)$ denote the lowest productivity of active firms, but this productivity does not depend on y^* . Instead, it is obtained by setting the value of the profit function, $\pi(q, 0, \varphi)$, equal to zero. In addition, the minimum productivity among bunched firms, denoted $\varphi^*(q, y^*)$, is the productivity at which $y(q, \varphi) = y^*$, whereas $\varphi^*(q, y^*) = \varphi^m(q, y^*)$ if no firms produce below y^* . In either case, total untaxed output may be written

$$Y^U = M \left(\int_{\varphi^m(q, y^*)}^{\varphi^*(q, y^*)} y(q, \varphi) f(\varphi) d\varphi + \int_{\varphi^*(q, y^*)}^{\varphi^{**}(q, p, b, y^*)} y^* f(\varphi) d\varphi \right), \quad (10)$$

where the first integral disappears if y^* is reduced to the point where no firm desires to produce below y^* . M is determined by the requirement that total supply equal demand:

$$Y^U + Y^T = X(q). \quad (11)$$

To fully demonstrate a missing middle, we must set up the government's optimal tax problem and characterize cases where it is indeed optimal for some firms to produce untaxed output. As discussed above, we first fix q and find the tax system that maximizes government revenue. The Lagrangian for this problem is:

$$L = M \left[\int_{\varphi^{**}(q, p, b, y^*)}^{\varphi^h} ((q - p)y(p, \varphi) + b - A) f(\varphi) d\varphi \right] \quad (12)$$

$$+ \lambda \left(X(q) - M \left[\int_{\varphi^m(q, y^*)}^{\varphi^*(q, y^*)} y(q, \varphi) f(\varphi) d\varphi + \int_{\varphi^*(q, y^*)}^{\varphi^{**}(q, p, b, y^*)} y^* f(\varphi) d\varphi + \int_{\varphi^{**}(q, p, b, y^*)}^{\varphi^h} y(p, \varphi) f(\varphi) d\varphi \right] \right)$$

$$+ \beta \left(\int_{\varphi^m(q, y^*)}^{\varphi^*(q, y^*)} \pi(q, 0, \varphi) f(\varphi) d\varphi + \int_{\varphi^*(q, y^*)}^{\varphi^{**}(q, p, b, y^*)} (qy^* - c(y^*, \varphi)) f(\varphi) d\varphi + \int_{\varphi^{**}(q, p, b, y^*)}^{\varphi^h} \pi(p, b, \varphi) f(\varphi) d\varphi - c_e \right).$$

The Lagrange multiplier, λ , multiplies the market-clearing constraint, and the Lagrange multiplier, β , multiplies the zero-profit constraint. These constraints account for the three types of firms: those below the cutoff (if there are any), which pay no taxes; those at y^* , which also pay no taxes; and those that produce in excess of the cutoff and are therefore taxed on their output. The control variables are the producer price p , which determines the tax $t = q - p$; the fee, b ; the number of entrants, M ; and the cutoff, y^* .

The zero-profit constraint has an important implication: because revenue $qX(q)$ must equal production costs plus tax payments, the maximization of tax payments net of administrative costs is equivalent to the minimization of total production costs plus administrative costs. To depict this minimization problem, we can write total production costs as a function of the number of taxed firms, $C(M^T)$, where the tax parameters are optimally chosen to achieve this number M^T ; then the optimal M^T minimizes $C(M^T) + AM^T$. This is an unconstrained maximization problem because the $C(M^T)$ reflects not only the adjustment to the producer price p to achieve zero profits (for the given consumer price q), but also the entry of firms to equate supply with demand. If a cutoff is desirable, it should be set so that raising it enough to reduce the number of taxed firms by a unit causes production costs to rise by an amount equal to the saving in administration costs: $-dC/dM^T = A$.

Note that production plus administrative costs are minimized only in a second-best sense, given the limitations on the available tax instruments. The second-best inefficiencies inherent in a cutoff rule will be studied below. If we could identify firm productivities and implement a productivity cutoff rather than an output cutoff, then $C+AM^T$ could be lower.

This minimization problem suggests a definition of deadweight loss from taxing any particular good that generates firm-level administrative costs: it is the excess of $C+AM^T$ over the minimum production costs, where the latter would be achieved in this competitive economy by setting y^* equal to zero and taxing all firms at the same rate. An important insight is that it is not generally optimal to simply minimize production costs when there are administrative costs of collecting taxes.

Returning to the revenue-minimization problem, the first-order condition for M shows that the Lagrange multiplier is the ratio of revenue, net of administrative costs, to output, which we denote T^e :

$$\lambda = \frac{R}{X} \equiv T^e. \quad (13)$$

The basic idea is that if there is an exogenous unit increase in output $X(q)$, then the number of firms in the industry can be increased to satisfy this output expansion, and the resulting increase in tax revenue, T^e , measures the social gain.

In the absence of a cutoff, T^e would equal the output tax rate, t , because the fixed fee b would equal administrative costs A , in its role as a Pigouvian tax. We show in the Appendix, however, that the optimal t exceeds T^e when firms do take advantage of a cutoff.²¹ Intuitively, the higher output tax, coupled with a lower fixed fee, makes relatively low levels of taxed output more attractive to firms, discouraging some of them from bunching at y^* . In other words, the fee is now no longer solely serving the role of a Pigouvian tax, but is also being used to control bunching. Given these conflicting considerations, we are unable to sign the difference between b and A when there is an output cutoff, although the excess of t over T^e limits the extent to which b can exceed A . Identifying this sign is not, however, needed for our subsequent results.

4. Missing Middle or Isolated Bottom?

The missing middle identified in Proposition 4 suggests an economy with a sizable number of small firms and large firms, but no intermediate-sized firms. But another possibility is that the level of the output cutoff used to exempt firms from taxation is so low that only a small number of firms take advantage of it, generating a size distribution of firms with intermediate- and large-sized firms and a few much smaller untaxed firms, representing an “isolated bottom” of the size distribution. The next proposition shows that this size distribution would never occur under an optimal output cutoff.

²¹ See Proposition A.1, which also shows that $b = A$ with heterogeneous firms and no output cutoff. Proposition A.2 states that $t > T^e$ under a sufficient condition that seems unlikely to be violated.

Proposition 5. *Starting from a welfare-maximizing tax system without output cutoffs, introducing a cutoff for firms in a given industry must lower welfare if the resulting set of untaxed firms is sufficiently small.*

The basic idea here is that if only a few firms can gain by choosing to produce untaxed output, then they cannot benefit much. Otherwise, there would be many other firms also benefiting from producing untaxed output. The benefits of the tax exemption for these few firms must be offset by the profit losses they incur to reduce their outputs to untaxed levels. In terms of Figure 1, if MC^h is the marginal cost curve for one of these firms, then the profit loss from reducing output to y^* approximately equals the benefit of the tax reduction once y^* is reached, $b + ty^*$. Thus, the movement of firms to untaxed output is not generating a rise in expected profits for the industry, and the government is therefore unable to raise taxes on firms above the cutoff without necessitating a rise in the consumer price to satisfy the zero-expected-profit requirement. But with some output now produced by untaxed firms and total output fixed at $X(q)$, there must be a reduction in total *taxed* output, even after we account for the entry of firms into the industry needed to keep supply equated with demand after some existing firms reduce their outputs to y^* . This decline in the tax base lowers tax revenue.²² As previously explained, welfare cannot be maximized if revenue is not maximized, given the consumer price q .

As the cutoff level is increased, more firms take advantage of it, and firms that are initially at the cutoff level see their profits rise as they increase output to the new higher cutoff and suffer a lower production disadvantage. Thus, the cutoff does start to raise expected profits for the industry, enabling the government to increase its taxes on firms above the cutoff without causing the consumer price to rise. Whether the higher taxes more than offset the loss in revenue from the exit of more firms from the tax base will depend on their tax payments relative to their administrative costs. If their tax payments do not exceed administrative costs much, then exempting them from taxation

²² We could eliminate this revenue loss by changing p and b in a way that eliminated incentives for firms to switch to untaxed output y^* . But this would violate the condition in the proposition that some firms do produce untaxed output.

will have little effect on the government budget, so the higher taxes on firms above the cutoff will result in a budget surplus, which can be used to raise welfare. Thus, we are able to prove:

Proposition 6. *For a given level of administrative costs, if the tax revenue collected from an industry in excess of these costs is sufficiently small in the absence of an output cutoff, then the welfare-maximizing tax system will involve a cutoff that is high enough to induce a positive measure of firms to produce untaxed output.*

Thus, the importance of administrative costs relative to tax collections is the critical determinant of whether there should be an output cutoff. Increasing the cutoff generates additional profits for existing untaxed firms, allowing the government to raise taxes on firms above the cutoff without necessitating a rise in the consumer price to keep expected profits equal to zero. But a higher cutoff reduces taxable output, which tends to reduce tax revenue. If the average net tax on output, T^e , is small enough, however, then this second effect will be unimportant, so additional revenue can be generated by raising the cutoff, at no cost to the consumer. This additional revenue can then be distributed to the consumer through a reduction in the consumer price.

Note that this welfare gain requires that the cutoff be high enough to induce a sizable number of firms to produce untaxed output. Consequently, Proposition 6 lends further support for the optimality of the “missing middle.” In particular, a welfare-improving output cutoff induces a sizable number of firms to opt out of the tax system by lowering their tax rates, so that there are many small untaxed firms as well as some large taxed firms, but there is a “missing middle” of intermediate-sized firms.

5. Nonlinear Tax Systems

Much of the potential welfare gain of an output cutoff for taxation may be offset by its negative impact on firms’ output decisions, as they attempt to qualify for tax-exempt status by reducing their outputs. One way to counteract these inefficiencies would be to give tax breaks to firms with taxable outputs near the cutoff. Simply stated, reducing the tax liability on those firms that might be tempted to cut their outputs to the cutoff level might keep them from so doing. This could be achieved using a nonlinear

tax system under which a net tax function, $T(y)$, is chosen so that firms choosing a relatively low output pay a relatively low average net tax, $T(y)/y$, where as before net means after administrative costs are subtracted. In this section, we provide a simple rule for taxing a firm that is indifferent between taxed and untaxed output, and we discuss possible shapes of the optimal nonlinear tax schedule. In particular, we argue that to keep the average tax rate relatively low at output levels near the cutoff, the marginal tax should be set so that taxed firms with higher outputs pay higher average net taxes than smaller taxed firms.

We again focus on a single industry and consider the sub-optimization problem of choosing the cutoff and nonlinear tax schedule to maximize tax revenue, given the chosen consumer price, q , which determines demand, $X(q)$, and welfare. The cutoff at y^* already gives us a special type of nonlinearity in the tax system, with $T(y) + A = 0$ for $y \leq y^*$; that is, there are no gross tax collections. But now we allow this aspect of the tax system's nonlinearity to be supplemented by marginal output taxes, dT/dy , that vary with income at $y > y^*$ rather than being set at a constant rate of t .

The formal treatment of the optimal nonlinear tax problem is in an additional appendix available on request. Here we provide a heuristic analysis based on Figure 2, which presents a possible tax schedule, beginning at y^* . A pre-tax profit indifference curve over tax payments and output is drawn for a type- φ^{**} firm, which maximizes profits at both y^* and y^{**} . We have constructed the tax schedule to rise above the type- φ^{**} firm's indifference curve as output increases from y^{**} , so that the firm will not choose any output between y^* and y^{**} . Firms with productivities between φ^{**} and some lower φ bunch at y^* .

This bunching could be eliminated by replacing the discontinuity in the tax schedule at y^* with a smooth tax schedule. But then there would exist some firms producing slightly above y^* and remitting almost no taxes, while the government incurred the administrative cost A to collect these taxes. Hence, the government would want to change the tax function so that those firms that do produce above y^* make tax payments that are large enough to justify the required administrative costs.

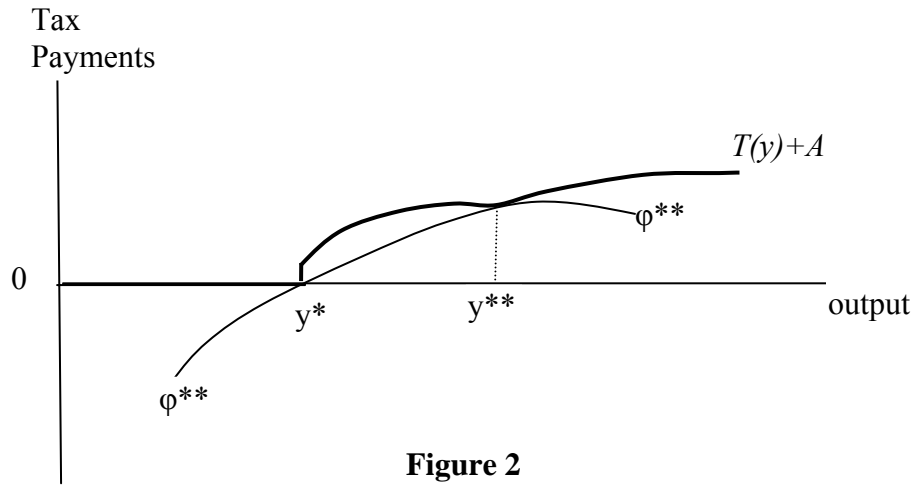


Figure 2

Consider now the rule for the optimal number of firms to exempt from taxation. The critical insight here is that this number can be increased by raising tax payments by a small amount in a small interval of incomes from y^{**} to $y^{**} + dy^{**}$. Because firms in this interval are approximately indifferent between y^* and y^{**} , this change causes a small number of firms to switch from taxed output to untaxed output, without significantly altering their profits. The direct revenue loss per firm is $T(y^{**})$, calculated net of the savings in administrative costs. Offsetting this loss, however, is the gain in revenue from an increase in entry into the industry. Since total demand stays fixed at $X(q)$, an existing firm's decision to reduce output from y^{**} to y^* must be offset by a change in the number of firms entering the industry, dM , that satisfies, $y^e dM = y^{**} - y^*$, where y^e is the output per firm entering the industry. Thus, $dM = \frac{y^{**} - y^*}{y^e}$. Following our previous notation,

let T^e be the average tax on output, calculated net of administrative costs, in which case output per firm is $T^e y^e$. Multiplying this amount by the change in M gives the total change in tax payments from this entry: $(y^{**} - y^*) T^e$. For the initial tax system to maximize revenue, this revenue gain must exactly offset the net loss in revenue, $T(y^{**})$, that directly resulted from the firm's switch from taxed output to untaxed output. Thus, the following condition must hold when the optimal number of firms is exempt from taxation:

$$T(y^{**}) = (y^{**} - y^*) T^e. \quad (14)$$

Condition (14) says that the nonlinear tax system is optimal only when there is no net change in tax revenue generated by a marginal firm's switch from taxed to untaxed output. An interesting aspect of this rule is that it contains no terms involving the behavioral responses of firms' outputs to tax changes. It is the rule that the government would want to follow to maximize tax payments if it had chosen an output cutoff y^* and could directly control the number of firms that produce at y^* or y^{**} . This direct control is effectively available to the government through its choice of the marginal firm's tax payments. In contrast, when the government is restricted to a linear tax system, it can alter the number of taxed firms only by changing the common marginal output tax rate t , which causes *all* taxed firms to alter their chosen output supplies.

In an additional appendix available on request, we derive tax schedules for firms above the cutoff y^{**} that possess the form shown in Figure 3, which depicts both the marginal and average net tax rates as functions of output. In particular, the average net tax for a firm with the lowest taxed output lies below the average net tax (i.e., $T(y^{**})/y^{**} < T^e$), a property that follows directly from (16). But the marginal tax on this output is high and declining, eventually falling below the average tax as the top output is approached.²³

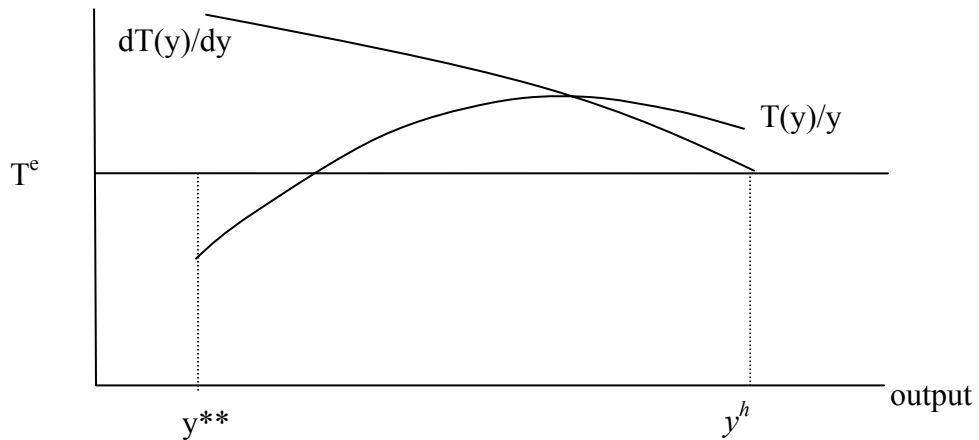


Figure 3

²³ Specifically, we show that $dT(y)/dy$ lies above T^e for $y^{**} \leq y < y(p, \phi^h)$, and declines to T^e at the top output level $y(p, \phi^h)$, if $(1-F(\phi))/f(\phi)$ goes to zero as ϕ goes to ϕ^h . This condition clearly holds for a uniform distribution. As a result, the average tax, $T(y)/y$, reaches a maximum at some y between y^{**} and y^h , as shown in Figure 3.

Although we have shown that allowing a nonlinear output tax can mitigate the distortions created in tax systems with an output cutoff rule, in practice this would clearly involve additional administrative costs. For this reason, the restriction to linear taxes should be taken seriously, and we next focus on that case when we derive a modified inverse elasticity rule. Nevertheless, the analysis in this section suggests the desirability of limited nonlinearities in the tax system, whereby the government couples the elimination of taxes at sufficiently low output levels with tax breaks for somewhat higher output.

6. A Modified Inverse-Elasticity Rule

When there are no untaxed firms producing in any industry, the standard inverse-elasticity rule derived in Proposition 2 holds. In that case, only the demand elasticity matters because raising the consumer price alone has no impact on the producer price received by firms, the level of which is fixed by the requirement that expected profits equal zero. With output cutoffs, in contrast, this rule no longer applies. Untaxed firms receive the consumer price, so supply elasticities matter. Thus, to state a new rule, we must first define the supply elasticity of untaxed output of good i :

$$\varepsilon_i^U = \frac{\partial Y_i^U}{\partial q_i} \frac{q_i}{Y_i^U} > 0. \quad (15)$$

For this definition, we hold fixed the *number* of firms producing untaxed output and consider only the marginal impact of the price they receive (i.e., the consumer price) on their output.

A change in q_i , and any accompanying changes in the other tax variables, will also generally change the share of firms that produce untaxed output, denoted F_i^{**} . We define this share elasticity as follows:

$$\varepsilon_i^F = \frac{dF_i^{**}}{dq_i} \frac{q_i}{F_i^{**}} > 0 \quad (16)$$

The change in this share directly affects the net revenue obtained from taxing good i . Letting R_i denote this net revenue, we define the revenue elasticity,

$$\varepsilon_i^R = \frac{dR_i}{dF_i^{**}} \frac{F_i^{**}}{R_i}. \quad (17)$$

The Appendix demonstrates that this revenue elasticity is negative (Proposition A.2).

These three elasticities all enter our modified inverse-elasticity rule. Again let α denote the consumer's marginal utility of income, and λ_B denote the marginal value of government revenue. With this notation, we now state the new rule as follows:

Proposition 7. *Under a linear tax system with optimal t_i and b_i for each good i and given output cutoffs, the average net tax rate for each taxed good satisfies the following modified inverse-elasticity rule.²⁴*

$$\frac{T_i^e}{q_i} = \frac{1 - \frac{\alpha}{\lambda_B}}{\varepsilon_i^X + \frac{Y_i^U}{X_i} \varepsilon_i^U - \varepsilon_i^R \varepsilon_i^F}. \quad (18)$$

This is a rule for the average net tax on output as a percentage of the consumer price — an average *ad valorem* net tax. Intuitively, it is the net tax rate that matters, not the gross rate, because a rise in output not only generates a social gain in the form of additional gross tax revenue, but also a social cost in the form of additional administrative costs. Recalling that $t_i > T_i^e$ (Proposition A.2 in the Appendix), this rule places lower bounds on the output tax rates.

The modified inverse-elasticity rule tells us that not only should, *ceteris paribus*, taxes be low on goods with high demand elasticities (as the Ramsey rule instructs) but, other things equal, they should also be low on goods with high supply elasticities for untaxed output. The basic idea is that taxing output at a higher rate distorts not only demand decisions, but it also distorts supply decisions by increasing untaxed output at the expense of taxed output. The supply elasticity reflects this latter distortion, because the

²⁴ We limit this rule to “taxed goods” because we have seen that not all goods are necessarily taxed in the presence of administrative costs.

positive impact of a higher consumer price on the supplies of existing untaxed firms crowds out taxed output, through a reduction in the number of firms entering the industry. But the higher consumer price reduces taxed output through a second avenue: firms producing taxed output find untaxed output more attractive, causing some of them to switch. The share elasticity accounts for this latter consideration, and its importance depends on the net revenue elasticity, which is shown in the Appendix to be negative (Proposition A): the movement of a marginal firm from taxed to untaxed output lowers revenue. Thus, goods with high share elasticities should, *ceteris paribus*, be taxed at relatively low net rates.

Note, finally, that a good with a relatively high share of output that is untaxed, given by $\frac{Y_i^U}{X_i}$ should have a relatively low average tax, all else equal. The reason is that the supply elasticity for untaxed output becomes more important in the modified inverse-elasticity rule as the untaxed-output share rises.

Now consider the case where the government uses a cutoff for some industries but not others. Without a cutoff, only demand elasticities enter the inverse-elasticity rule. With a cutoff, the supply-related elasticities also enter, and they all contribute to a lower tax rate. Thus, we have:

Proposition 8. *Assume that t_i and b_i are optimal for each good i , given exogenous output cutoffs. If taxed goods i and j have the same demand elasticity, but the government uses a cutoff rule only for i , then the optimal average net tax on output is lower for i than for j :*

$$\frac{T_i^e}{q_i} < \frac{T_j^e}{q_j} .$$

Presumably, a cutoff rule is more likely to be used in industries with relatively high administrative costs. If this is the case, Proposition 8 tells us that the industries with the high administrative costs tend to have the lower average net taxes. In other words, any additional gross tax payments are not fully covering the higher administrative costs. The intuition is that, although a cutoff rule lowers the total cost of taxing a sector's

firms, it increases the marginal cost of so doing because a higher tax rate causes both inefficient supply and demand responses.

Total administrative costs will tend to be relatively large for industries whose technologies dictate that they will consist of relatively small firms. To the extent that the government responds by using a cutoff rule for industries populated by small firms, but not those with large firms, Proposition 8 also suggests that low taxes should be levied on small firms, net of these administrative costs. Of course, their gross taxes would be high enough to cover these administrative costs.

Return now to the case where a nonlinear tax on output is possible. While the inverse elasticity rule also holds in this case, it is possible to use (14) to show that the revenue elasticity equals zero.²⁵ Thus, the rule reduces to

$$\frac{T_i^e}{q_i} = \frac{1 - \frac{\alpha}{\lambda_B}}{\varepsilon_i^X + \frac{Y_i^U}{X_i} \varepsilon_i^U} \quad (19)$$

Comparing (19) with (18), we see that if the government uses nonlinear taxes for some, but not all, goods, then, all else equal, goods facing nonlinear taxes should be taxed more heavily in terms of average net tax rates than goods facing linear taxes. This is not surprising, since the use of nonlinear taxes must reduce the distorting effects of raising revenue.

Finally, it is interesting to observe that (19) would also hold in the case where the government could exempt firms from taxation based on their productivities, rather than observations of output. There would be no “share elasticity” in this case, because firms would not be able to switch their tax status in response to a change in tax rates. This does not arise in the tax systems studied in this article because of the presumption that a firm’s productivity is not observable by the tax authority.

7. Conclusions

To be relevant to a world (like ours) in which there are significant administrative costs to collecting taxes, a theory of optimal tax systems must address who or what

²⁵ Combine (14) with (A14) in the Appendix. Establishing this result does not actually require that the entire nonlinear tax schedule be optimal. Rather, we are using only the optimality rules for firms at the margin between taxed and untaxed output

entities remit tax as well as what triggers a tax. This requires attention to the role of firms in tax systems, and to the difficult problem of collecting tax from small firms.

We develop models that produce some insight into the optimal design of tax systems when there are significant fixed per-firm costs of collection, and demonstrate that under some conditions the optimal tax system will generate a missing middle of firm sizes. These models extend the insights of Yitzhaki (1979) and Wilson (1989), who assume that there is an exogenous distribution of collection costs across goods, and that a good may be either completely exempt from taxation or taxed uniformly, and of Keen and Mintz (2004), who derive an optimal cutoff rule for a value-added tax in a stylized setting. We provide an explanation for varying collection costs based on differing technologies that generate different size distributions of firms across industries. These differences, in turn, render more or less important the fixed per-firm collection costs.

In this setting, we generalize the policy instruments available to the tax authority by allowing them to collect taxes from some, but not all, firms in an industry. In addition, we show that a fixed per-firm fee can be an important component of an optimal tax system, and we show that average net (of administrative costs) taxes on output should differ across industries, depending not only on the elasticity of demand for the good (as in a standard Ramsey model), but also on the size distribution of firms and the supply responses of firms to a tax increase.

The models we have developed in this paper also show that optimal remittance systems generally induce production inefficiency. This is in contrast to the well-known finding of Diamond and Mirrlees (1971) that, under certain assumptions including the absence of administrative costs, an optimal tax system will always satisfy aggregate production efficiency. This suggests that optimal industrial organization must be considered together with optimal tax policy.

Future modeling work might usefully extend some aspects of this model that are highly stylized. For example, the models in this paper presume that all goods are produced in a single stage of production and that each firm produces only one good. Thus, they cannot address the important production efficiency questions that arise in the analysis of cascading business turnover taxes (also known as gross receipts taxes). Nor can the present model address without some refinement an important issue that arises in

the implementation of a value-added tax: the exemption of firms that sell to non-exempt firms (or to exempt firms that sell to non-exempt firms) does not truly relieve these firms of some tax burden, unless it induces the formation of “chains” of exempt firms selling to each other and ultimately to final consumers.²⁶ It would be useful to allow the administrative cost to be itself subject to policy control, as stressed by Slemrod and Yitzhaki (2002).²⁷ Finally, taxation is not the only form of government intervention in the production process, and many other types of regulation (such as health and safety standards, or antidiscrimination laws) explicitly or implicitly exempt small firms, creating many of the same incentives addressed in this paper, although without the same direct implications for revenue.²⁸ Ideally, models of government policy toward small firms should address all aspects of taxation and regulation simultaneously.²⁹ Because of the ubiquitous reliance of real tax systems on remittance by firms, optimal tax models that are, in the words of Hahn (1973), “guides to action,” must feature heterogeneous firms and costs of collection that are regressive with respect to firm size. We offer this set of models as a first step in the development of such a framework, one that would apply generally to the environments in which tax systems of both developed and developing countries operate.

²⁶On the formation of VAT chains, see de Paula and Scheinkman (2007).

²⁷ For example, Gordon and Li (2005) argue that when businesses and individuals have regular dealings with financial institutions, the cost of monitoring their tax affairs falls, and policy should effectively subsidize these interactions.

²⁸ For example, this framework could potentially be used to analyze the optimal threshold for the implementation of a regulatory policy such as the requirement of accommodations under the Americans with Disabilities Act (ADA). Because of the per-firm administrative costs associated with government monitoring of compliance by small businesses (or the costs to the court system of extensive litigation relating to small businesses), it may be optimal to exempt firms below a certain size from the ADA. The resulting production inefficiencies (and the lower social benefits from limiting the scope of the ADA) would be traded off against savings in administrative costs.

²⁹ Krueger (2007) reports that, since 1950, in India firms employing less than 10 workers have been exempt from many regulations governing employment of workers, provision of pensions and other safety net items. In response, many larger firms do not register as required, and in other cases factories “have office doors with different names on each one in order to keep under the limit of ten!” (p. 24). The latter behavior would be an example of misrepresenting firm size, which is not addressed by the models of this paper. The reorganization of firms solely for tax purposes may, though, affect the measured size distribution of firms even if it does not affect actual production operations; see Sivadasan and Slemrod (2007) for an example in which a change in the tax treatment of partnerships in India distorted some income inequality measures based on firm-provided data. Note finally, as Tybout (2000) stresses, that in many countries tax and regulatory policies favor large over small companies.

Appendix

Proof of Proposition 1. Let $b_i > A_i$. Then lower b_i slightly, and offset the gain in profits, $M_i db_i$ by lowering p_i with q_i fixed (i.e., by levying a higher t_i): $-M_i db_i = X_i dt_i$. Then the zero-profit requirement remains satisfied. If there were no behavioral changes, we would then see that tax revenue stays the same. Demand X_i stays the same, since it is determined by q_i , which has not changed; but the fall in p_i lowers output per firm. As a result, more firms must enter the industry to keep total output equal to the fixed demand. None of these behavioral changes affect the zero-profit requirement, but revenue rises by $(b_i - A_i)dM_i > 0$. Thus, we have a revenue gain, which corresponds to a welfare gain because the surplus in the government budget can be used, e.g., to lower q_i . By reversing the argument, we find that b_i cannot be less than A_i . Thus, $b_i = A_i$. Q.E.D.

Proposition A.1. *In the model with heterogeneous firms but no output cutoff, the optimal linear tax system satisfies $b = A$.*

Proof of Proposition A.1. Let $b > A$ initially, and repeat the previous proof by lowering b and raising the tax t (with q fixed) so that expected profits remain equal to zero: $-M(1-F^{**})db = Xdt$, where F^{**} equals the fraction of firms entering the industry that choose to exit after learning their productivities. If there were no behavioral changes, tax revenue would stay fixed. Total demand again stays fixed, but the fall in the producer price, $p = q - t$, lowers output per firm, which by itself causes more firms to enter the industry to keep total output equal to the fixed demand: M rises.

A new consideration here is that firms that are initially indifferent between producing and exiting the industry now strictly prefer to produce, because their outputs are relatively low, reducing the burden of the higher output tax: $-db = [X/(M(1-F^{**}))]dt > ydt$ for firms producing y below average output of producing firms, $X/(M(1-F^{**}))$. As a result, the minimum output of firms producing in the industry declines. In other words, there is a decline in the share of firms that exit the industry, denoted F^{**} . This consideration also lowers the average output per firm, necessitating a further rise in the number of firms that produce in the industry to keep total output equal to the fixed demand: $M(1 - F^{**})$ rises.

Thus, the policy perturbation raises $M(1 - F^{**})$, causing tax revenue to rise by $(b-A)d(M(1 - F^{**}))$. As before, this revenue gain can be used to raise welfare, implying that it could not have been optimal to set $b > A$. Similarly, the optimal b cannot be less than A . Q.E.D.

Proof of Proposition 2. Differentiating the Lagrangian with respect to M_i , we obtain a first-order condition that implies

$$\frac{\lambda_i}{\lambda_B} = \frac{R_i}{X_i} = t_i, \quad (\text{A1})$$

where R_i is the amount of revenue raised by taxes on firms in sector i , calculated net of administrative costs. Revenue per unit of output equals the output tax because administrative costs are financed by the fixed fee (Proposition 1). By differentiating the Lagrangian with respect to q_i and employing this equality, we obtain a first-order condition that implies (4). Q.E.D.

Proof of Proposition 3. Consider eliminating the output tax on some good j , causing a loss of net tax revenue equal to $t_j X_j(q_j)$. Then the consumer price drops from q_j by the amount t_j to the producer price p_j (see Figure A). The gain in consumers' surplus exceeds the lost tax revenue:

$$\Delta_i CS_j \equiv \int_{p_j}^{q_j} X_j(p_j) dp_j = t_j X_j(q_j) \left(1 + .5 \frac{t_j}{q_j} \varepsilon_j^X \right), \quad (\text{A2})$$

which is equal to tax revenue plus the deadweight loss from the tax, using the quadratic expression for the deadweight loss. Eliminating the fixed fee ($b_i = A_i$) results in a further price drop to a lower consumer price, q_j^* , to keep profits equal to zero, where q_j^* is implicitly defined by:

$$M_j \int_{q_j^*}^{p_j} y_j(p_j) dp_j = M_j A_j. \quad (\text{A3})$$

Equation (A3) gives

$$(p_j - q_j^*)y_j(p_j^\#) = A_j \quad (\text{A4})$$

for some $p_j^\#$ about halfway between p_j and q_j^* . The increase in consumers' surplus from the elimination of the fixed fee resulting from this price reduction satisfies

$$\begin{aligned} \Delta_a CS &> (p_j - q_j^*)X_j(p_j) = \frac{M_j A_j}{M_j y_j(p_j^\#)} X_j(p_j) \\ &= \frac{M_j A_j}{M_j y_j(p_j^\#)} (X_j(q_j) - t_j X_j'(q_j)) > M_j A_j \left(1 + \frac{t_j}{q_j} \varepsilon_j^X \right). \end{aligned} \quad (\text{A5})$$

The first inequality holds because the change in consumers' surplus includes the expansion in output beyond $X_j(p_j)$ (see Figure A). The first equality uses (A4), the second equality uses a linear approximation, and the last inequality holds because $M_j y_j(p_j^\#) < M_j y_j(p_j) = X_j(q_j)$.

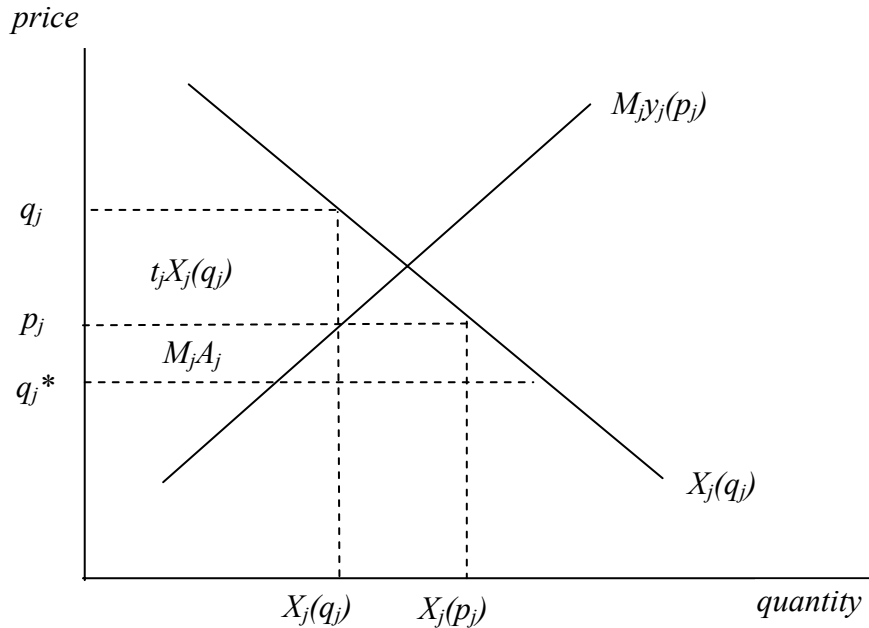


Figure A

On the other hand, only the lost revenue from the output tax needs to be recouped through a rise in the taxes on the remaining goods; the decline in fixed fees is exactly offset by a decline in administrative costs. Assuming this revenue loss is small relative to total tax revenue, it will not matter which tax rates on other goods are increased to offset the revenue loss, as long as these tax rate increases are kept small rather than being focused on a small group of other goods (in which case first-order approximations are not appropriate because the tax changes are large). Thus, the loss in consumers' surplus from raising another dollar of revenue is:

$$\frac{dCS_i}{dR_i} = \frac{X_i}{X_i + t_i(dX_i/dq_i)} = \frac{1}{1 - \frac{t_i}{q_i} \varepsilon_i^X}, \quad (\text{A6})$$

where i is any taxed good other than j . This expression is the marginal cost of public funds, denoted $MCPF$ in the text.

Multiplying $MCPF$ by the lost net revenue, $t_j X_j(q_j)$, and comparing to the increase in consumers' surplus given by the sum of (A2) and (A5) then gives the following sufficient condition for an increase in welfare:

$$1 + .5 \frac{t_j}{q_j} \varepsilon_j^X + \frac{M_j A_j \left(1 + \frac{t_j}{q_i} \varepsilon_j^X \right)}{t_j X_j(q_j)} > \frac{1}{1 - \frac{t_i}{q_i} \varepsilon_i^X}, \quad (\text{A7})$$

which becomes

$$\left(1 + .5 \frac{t_j}{q_j} \varepsilon_j^X \right) \left(1 + \frac{M_j A_j}{t_j X_j(q_j)} \right) > \frac{1}{1 - \frac{t_i}{q_i} \varepsilon_i^X}, \quad (\text{A8})$$

and then

$$\frac{M_j A_j}{t_j X_j(q_j)} > \frac{1}{\left(1 - \frac{t_i}{q_i} \varepsilon_i^X \right) \left(1 + .5 \frac{t_j}{q_j} \varepsilon_j^X \right)} - 1. \quad (\text{A9})$$

Using the definition of the *MCPF* given by (A6), we may manipulate the right side of (A9) into the expression given in Proposition 3. Q.E.D.

Proof of Proposition 4. This proposition clearly holds if $b + ty^* > 0$, because any firm producing output slightly above y^* would benefit from lowering output by the small amount needed to eliminate its tax burden. Thus, we need only show that this inequality holds. Suppose instead that $b + ty^* = 0$. Then, because marginal costs vary continuously with productivity, the presence of some firms at y^* implies that there will be other firms producing outputs slightly above y^* and paying almost no taxes. But the government would be incurring the administrative cost A to collect a negligible amount of taxes. By raising y^* it would incur almost no revenue loss, while eliminating these administrative costs. Thus, the original y^* could not have been optimal. A similar argument shows that we cannot have $b + ty^* < 0$. Q.E.D.

Proof of Proposition 5. If no firms produced untaxed output, then we know that revenue is maximized by setting $b = A$. Given this equality, raise y^* to the highest point where there is no bunching ($\varphi^{**} = \varphi^*$). Then only type- φ^{**} firms are willing to produce untaxed output, and these firms are just indifferent to reducing their output to qualify for tax-exempt status. To induce a marginal number of firms to move to untaxed output, we may then either raise b or y^* or both; the welfare effect will be the same, given that b and p are initially optimized for the case of untaxed firms. Let us differentiate the Lagrangian given by (12) for the revenue-maximization problem with respect to y^* :

$$\frac{\partial L}{\partial y^*} = T^e (y^{**} - y^*) \frac{\partial F^{**}}{\partial y^*} - ty^{**} \frac{\partial F^{**}}{\partial y^*} = -ty^* \frac{\partial F^{**}}{\partial y^*} < 0, \quad (\text{A10})$$

where $T^e = t$ when $b = A$ and use is made of (13) and $\partial F^{**}/\partial y^*$ is defined as the marginal rise in the share of firms that produce untaxed output. Thus, starting from the highest cutoff at which there are no untaxed firms, inducing some firms to choose untaxed output creates a first-order revenue loss for a fixed value of q , implying that welfare must fall. Q.E.D.

Proof of Proposition 6. Returning to the Lagrangian given by (12) for the revenue-maximization problem, set $b = A$ (as required by Proposition 1 in the absence of a cutoff) and $t = 0$, and differentiate the Lagrangian with respect to y^* :

$$\frac{\partial L}{\partial y^*} = \beta \int_{\varphi^*}^{\varphi^{**}} (q - c_y(y^*, \varphi)) f(\varphi) d\varphi, \quad (\text{A11})$$

where use is made of (13). As y^* increases above the highest level where no firm strictly benefits from producing y^* ($\varphi^{**} = \varphi^*$), firms bunch at y^* ($\varphi^{**} > \varphi^*$) and further increases in y^* raise profits for these bunched firms, because they sell at a price q above their marginal costs at y^* . Thus, the expected profits available to a firm entering the industry rise. Equation (A11) measures the marginal value of the profits generated by a unit rise in y^* . By raising y^* to generate these profits, the government can then satisfy the zero-expected-profit requirement by raising the tax t and fee b . Thus, government revenue rises, implying a welfare gain. Because revenue rises when $t = 0$ at the no-cutoff optimum, the continuity properties of the model imply that revenue will also rise when this t is positive but sufficiently small. Q.E.D.

Proof of Proposition 7. To derive the new rule, recall the previous sub-optimization problem: given a good's consumer price, the government maximizes the revenue obtained from taxing the good for given output and zero profits. Let $R_i(q_i)$ denote this maximized value of revenue for good i . This function will depend on whether linear or nonlinear tax schedules are used, and on whether a cutoff rule is used. The optimal tax problem then consists of maximizing the indirect utility function, subject to the government budget constraint:

$$\begin{aligned} \text{Max } & \Sigma_i v_i(q_i) \\ \text{s.t. } & \Sigma_i R_i(q_i) = E, \end{aligned}$$

where E is, as before, the government's revenue requirement. Using Roy's Identity, the first-order condition for q_i is

$$\frac{\alpha}{\lambda_B} X_i(q_i) = \frac{dR_i(q_i)}{dq_i}, \quad (\text{A12})$$

where α is the consumer's marginal utility of income, and λ_B is the Lagrange multiplier on the government's budget constraint, or the marginal social value of government revenue. To obtain a modified inverse-elasticity rule, we therefore need to calculate the revenue derivative. By the envelope theorem, this derivative is simply the derivative of the Lagrangian for the revenue-maximization problem. Also according to the envelope theorem, this derivative will not depend on whether we also change the fixed fee or the producer price as the consumer price rises, because both have a zero marginal impact on the Lagrangian at the optimum.

Omitting subscripts to simplify notation, let us then raise q while also increasing the fixed fee by an amount, db/dq , that keeps expected profits equal to zero: $db/dq = Y^U/(M(1-F^{**}))$. Differentiating the Lagrangian given by (12) with respect to q and this change in b , and using the previously-derived equality, $\lambda = T^e$, gives

$$\frac{dR}{dq} = X + T^e \frac{dX}{dq} - T^{**} M \frac{dF^{**}}{dq} - T^e \left(\frac{\partial Y^U}{\partial q} - (y^{**} - y^*) M \frac{dF^{**}}{dq} \right), \quad (\text{A13})$$

where T^{**} denotes the net tax payments for the type- φ^{**} firm (i.e., $T^{**} = ty^{**} + b - A$) and $\frac{dF^{**}}{dq}$ is the derivative of the share of firms that are untaxed with respect to q and the "profit-preserving" rise in b .

Changing the share of firms that produce untaxed output will alter tax revenue, R_i , by

$$\frac{dR_i}{dF_i^{**}} = M_i \left(T_i^e (y_i^{**} - y_i^*) - T_i^{**} \right). \quad (\text{A14})$$

In expression (A14), T_i^{**} is the net revenue obtained from a marginal taxed firm, which is now lost as some of these firms move from output y_i^{**} to untaxed output y_i^* , and $T_i^e (y_i^{**} - y_i^*)$ is the resulting rise in revenue from the resulting entry of additional firms to increase total output back to total demand to offset the lower per-firm output of

untaxed firms. In the previous section we proved that this derivative is equal to zero under optimal nonlinear taxation. But, in the case of linear taxation, we shall show below that revenue falls.

Using (A14), we may rewrite (A13) as

$$\frac{dR}{dq} = X + T^e \left(\frac{dX}{dq} - \frac{\partial Y^U}{\partial q} \right) + \frac{dR}{dF^{**}} \frac{dF^{**}}{dq}. \quad (\text{A15})$$

Substituting this derivative into (A32) and expressing the result in elasticity form yields (18) in Proposition 7. Q.E.D.

Proposition A.2. *If some firms in a given industry take advantage of an output cutoff, and if output per firm in taxed industry i is at least as great as $y_i^{**} - y_i^*$, then the optimal taxation of that industry satisfies $t_i > T_i^e$ and $\varepsilon_i^R < 0$.*

Proof of Proposition A.2. Again omitting subscripts, hold q fixed and raise p (i.e., lower t) while increasing the fixed fee b to again keep expected profits fixed. For the initial p and b to be optimal, the resulting change in the Lagrangian given by (12) must equal zero at the optimum:

$$\begin{aligned} \frac{\partial L}{\partial p} + \frac{\partial L}{\partial b} \frac{db}{dp} = T^e M \left[(y^{**} - y^*) \frac{dF^{**}}{dp} - \int_{\varphi^{**}}^{\varphi^h} \frac{\partial y(p, \varphi)}{\partial p} f(\varphi) d\varphi \right] \\ + M \left[t \int_{\varphi^{**}}^{\varphi^h} \frac{\partial y(p, \varphi)}{\partial p} f(\varphi) d\varphi - (ty^{**} + b - A) \frac{dF^{**}}{dp} \right] = 0. \end{aligned} \quad (\text{A16})$$

The change in the share of firms that are untaxed, dF^{**}/dp , is positive because although expected profits have not changed, the rise in p and b have increased the tax burden of type- φ^{**} firms, since their relatively low output y^{**} implies that they benefit relatively little from the higher p .

Suppose that

$$t \leq T^e \equiv t \frac{Y^T}{X} + \frac{(b - A)M(1 - F^{**})}{X}. \quad (\text{A17})$$

Then (A16) gives

$$T^e M \left[(y^{**} - y^*) \frac{dF^{**}}{dp} \right] - M \left[(ty^{**} + b - A) \frac{dF^{**}}{dp} \right] \geq 0. \quad (\text{A18})$$

Substituting the expression for T^e given by (A17) into (A18) yields

$$t \left(\left(\frac{Y^T}{X} \right) (y^{**} - y^*) - y^{**} \right) + (b - A) \left(\frac{M(1 - F^{**})(y^{**} - y^*)}{X} - 1 \right) \geq 0. \quad (\text{A19})$$

The first of the two terms in (A19) is clearly negative. Since (A17) also implies that $b > A$, the assumption that $(y^{**} - y^*) < X/[M(1 - F^{**})]$ implies that the second term is non-positive. Thus, (A18) is violated, allowing us to conclude that $t > T^e$. Then (A36) implies the remaining part of the proposition. Q.E.D.

References

- Auerbach, Alan and James R. Hines. 2002. "Taxation and Economic Efficiency." In A. Auerbach and M. Feldstein (eds.), *Handbook of Public Economics*, Volume 3, North-Holland.
- Auriol, Emmanuelle and Michael Warlters. 2005. "Taxation Base in Developing Countries." *Journal of Public Economics*, 89(4): 625-646.
- Ballard, Charles. 1997. "Marginal Efficiency Cost Calculations for Different Types of Government Expenditure: A Review." In John G. Head and Richard Krever (eds.), *Taxation Towards 2000*, Melbourne: Australian Tax Research Foundation: 537-563.
- Christensen, K., Robert Cline, and Thomas Neubig. 2001. "Total Corporate Taxes: "Hidden," Above-the-Line, and Non-Income Taxes." *National Tax Journal* 54(3) (September): 495-506.
- De Paula, Áureo and José Scheinkman. 2007. "The Informal Sector." Working paper, University of Pennsylvania, August.
- Diamond, Peter and James Mirrlees. 1971. "Optimal Taxation and Public Production, Part I: Production Efficiency." *American Economic Review* 61 (March): 8-27.
- Djankov, Simeon, Rafael La Porta, Florencio López de Silanes and Andrei Shleifer. 2002. "The Regulation of Entry." *Quarterly Journal of Economics* 117(1): 1-37.
- Fortin, Bernard, Nicholas Marceau and Luc Savard. 1997. "Taxation, Wage Controls and the Informal Sector." *Journal of Public Economics*, 66(2): 293-312.
- Friedman, Eric, Simon Johnson, Daniel Kauffmann and Pablo Zoido-Lobaton. 2000. "Dodging the Grabbing Hand: The Determinants of Unofficial Activity in 69 Countries." *Journal of Public Economics*, 76(3): 459-493.
- Gauthier, Bernard and Mark Gersovitz. 1997. "Revenue Erosion through Tax Exemption and Evasion in Poor Countries." *Journal of Public Economics* 64(3): 404-424.
- Gordon, Roger H. and Wei Li. 2005. "Tax Structure in Developing Countries: Many Puzzles and a Possible Explanation." NBER Working Paper No. 11267, April.
- Hahn, Frank. 1973. "On Optimum Taxation." *Journal of Economic Theory* 6(1): 96-106.
- Heller, Walter P. and Karl Shell. 1974. "On Optimal Taxation with Costly Administration." *American Economic Review Papers and Proceedings* 64 (May): 338-345.
- Hopenhayn, Hugo A. 1992a. "Entry, Exit, and Firm Dynamics in Long Run Equilibrium." *Econometrica* 60(5) (September): 1127-1150.

Hopenhayn, Hugo A. 1992b. "Exit, Selection, and the Value of Firms." *Journal of Economic Dynamics and Control* 16(3-4): 621-653.

International Tax Dialogue (with input from the staff of the International Monetary Fund, Inter-American Development Bank, OECD, and the World Bank). 2007. *Taxation of Small and Medium Enterprises*. Background paper for the International Tax Dialogue Conference, Buenos Aires. October.

Keen, Michael and Jack Mintz. 2004. "The Optimal Threshold for a Value-Added Tax." *Journal of Public Economics* 88(3) (March): 559-576.

Kopczuk, Wojciech. 2003. "A Note on Optimal Taxation in the Presence of Externalities." *Economics Letters* 80(11) (July): 81-86.

Kopczuk, Wojciech and Joel Slemrod. 2006. "Putting Firms into Optimal Tax Theory." *American Economic Review Papers and Proceedings* 96 (May): 130-134.

Krueger, Anne. 2007. "The Missing Middle." Presented at Stanford Institute of Economic Policy Research conference on India, June 7-9.

Melitz, Marc J. 2003. "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity." *Econometrica* 71(6): 1696-1725.

Mirrlees, James. 1971. "An Exploration in the Theory of Optimum Income Taxation." *Review of Economic Studies* 38 (April): 175-208.

Rauch, James E. 1991. "Modeling the Informal Sector Formally." *Journal of Development Economics* 35: 33-47.

Sandmo, Agnar. 1975. "Optimal Taxation in the Presence of Externalities." *The Swedish Journal of Economics* 77(1): 86-98.

Sivadasan, Jagadeesh and Joel Slemrod. Forthcoming. "Tax Law Changes, Income Shifting, and Measured Wage Inequality: Evidence from India." *Journal of Public Economics*.

Slemrod, Joel. 2006. "The Compliance Cost of Taxing Business." Mimeo. University of Michigan.

Slemrod, Joel. Forthcoming. "The Economics of Tax Remittance." *National Tax Journal*.

Slemrod, Joel. 2008. "Optimal Taxation with Tax Enforcement Externalities." In process. University of Michigan.

Slemrod, Joel and Wojciech Kopczuk. 2002. "The Optimal Elasticity of Taxable Income." *Journal of Public Economics* 84(1): 91-112.

Slemrod, Joel, John Whiting, and Jonathan Shaw. 2007. "Tax Implementation Issues in the United Kingdom." Prepared for the "Mirrlees Committee" on tax reform in the U.K., sponsored by the Institute for Fiscal Studies, London.

Slemrod, Joel and Shlomo Yitzhaki. 2002. "Tax Avoidance, Evasion, and Administration." In Alan Auerbach and Martin Feldstein, (eds), *Handbook of Public Economics*. Amsterdam, London, and New York: Elsevier.

Tybout, James R. 2000. "Manufacturing Firms in Developing Countries: How Well Do They Do, and Why?" *Journal of Economic Literature* 38(1): 11-44.

Wilson, John D. 1989. "On the Optimal Tax Base for Commodity Taxation." *American Economic Review*. 79(5) (December): 1196-1206.

Yitzhaki, Shlomo. 1979. "A Note on Optimal Taxation and Administrative Cost." *American Economic Review* 69 (June): 475-480.

Zee, Howell. 2005. "Simple Analytics of Setting the Optimal VAT Exemption Threshold." *De Economist* 153(4): 461-471.